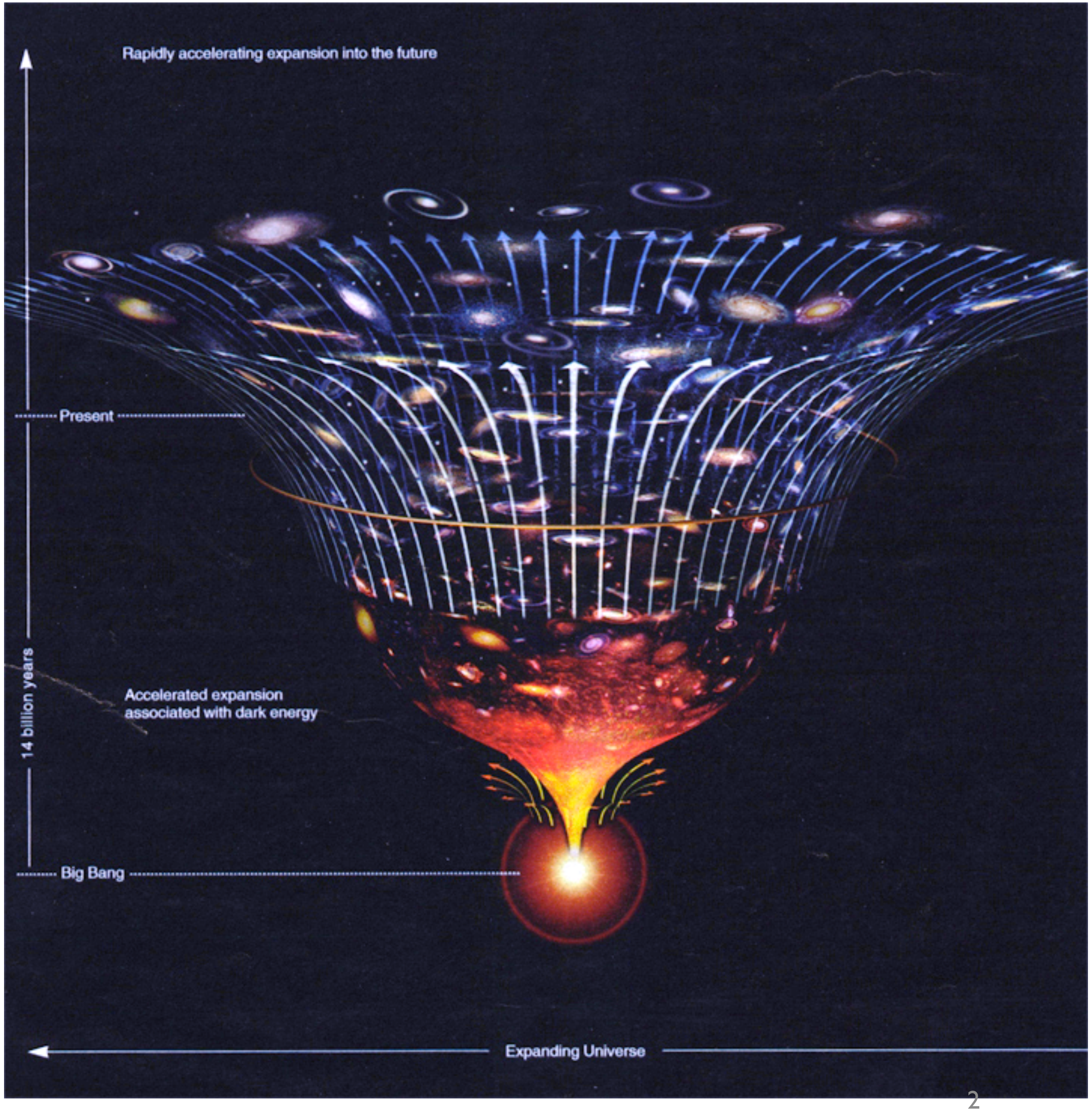
Using Manifold Learning to Maximize the Information from Cosmology Surveys of the 2020s

Daniel Masters, NPP fellow

Jet Propulsion Laboratory / California Institute of Technology

Postdoc Seminar Series

May 30, 2019



Breakthrough from the 1990s:
Accelerating cosmic expansion

2011 Nobel Prize in Physics

Λ?

Tension of local H_0 measurement with CMB-based value now at 4.4σ (Riess et al. 2019)

Dark Energy

Dark Energy affects the:

Expansion history of the Universe

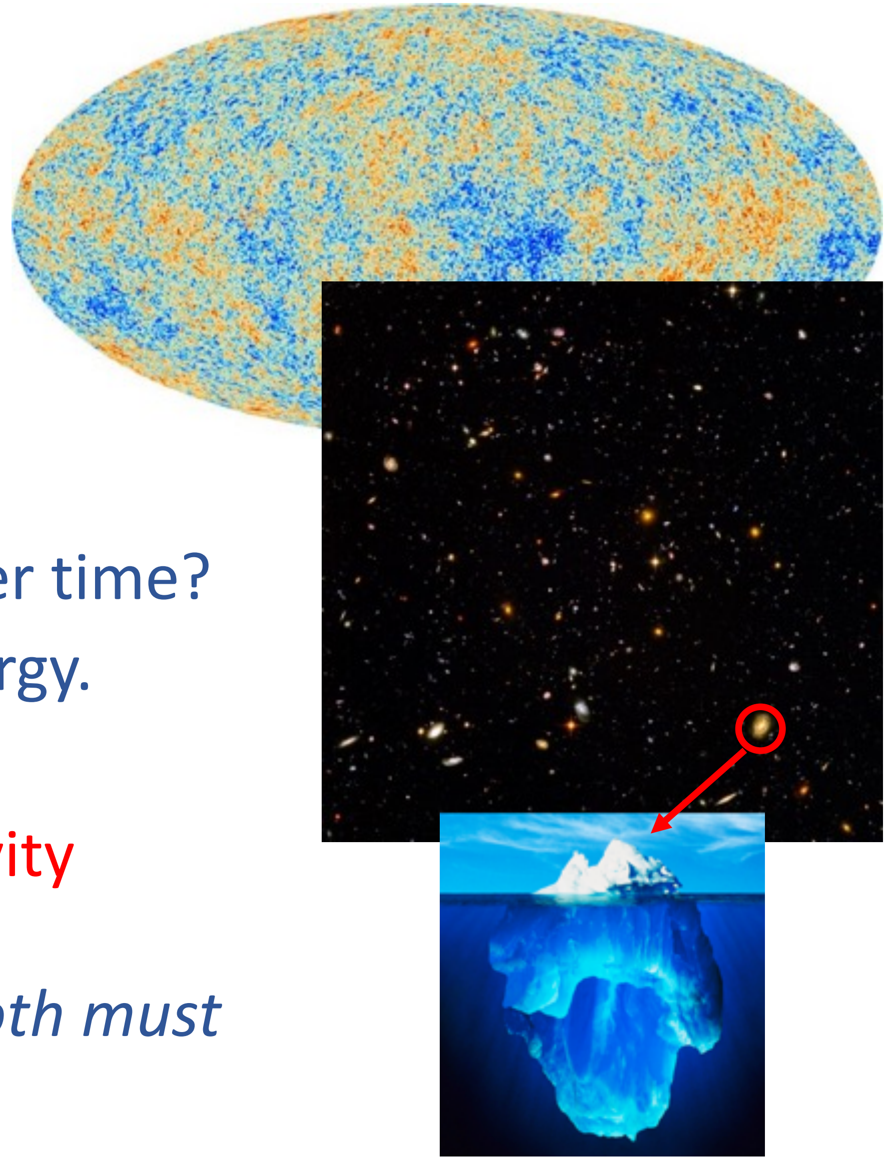
- How fast did the Universe expand?
- Also known as the geometry of the Universe.

Growth of structures

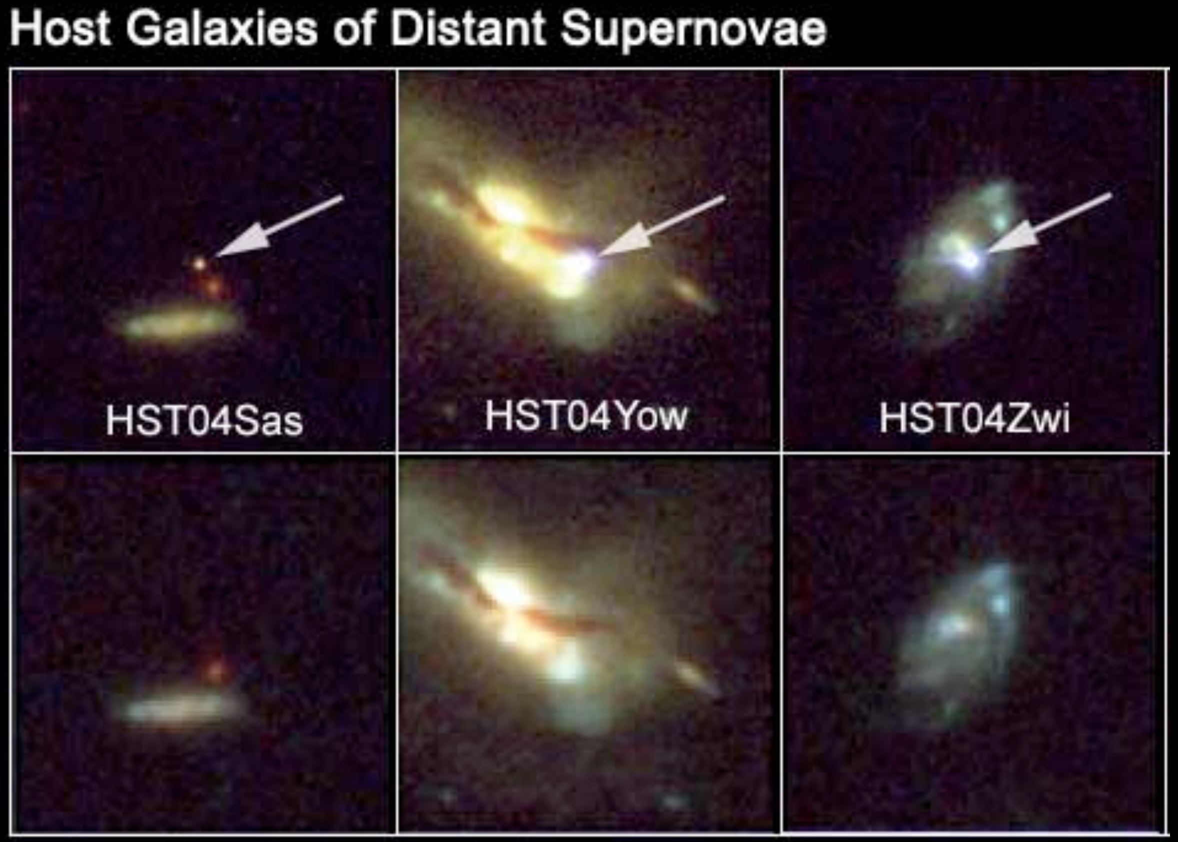
- How do dark matter structures evolve and grow over time?
- Attractive gravity competes with repulsive dark energy.

If Einstein’s General Relativity is wrong, modified gravity theories could explain the accelerating expansion.

This would change the effects above differently, so *both must be measured*



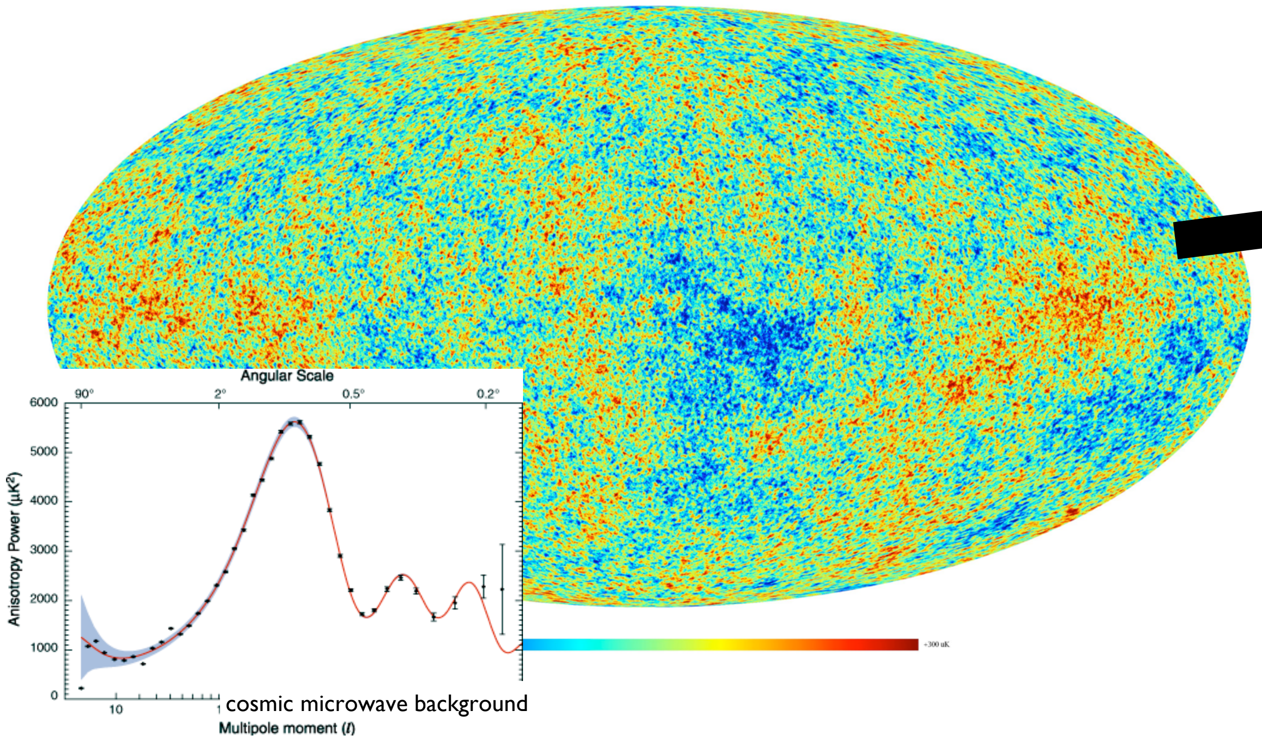
Probes of the Cosmic Expansion History: Standard Candles



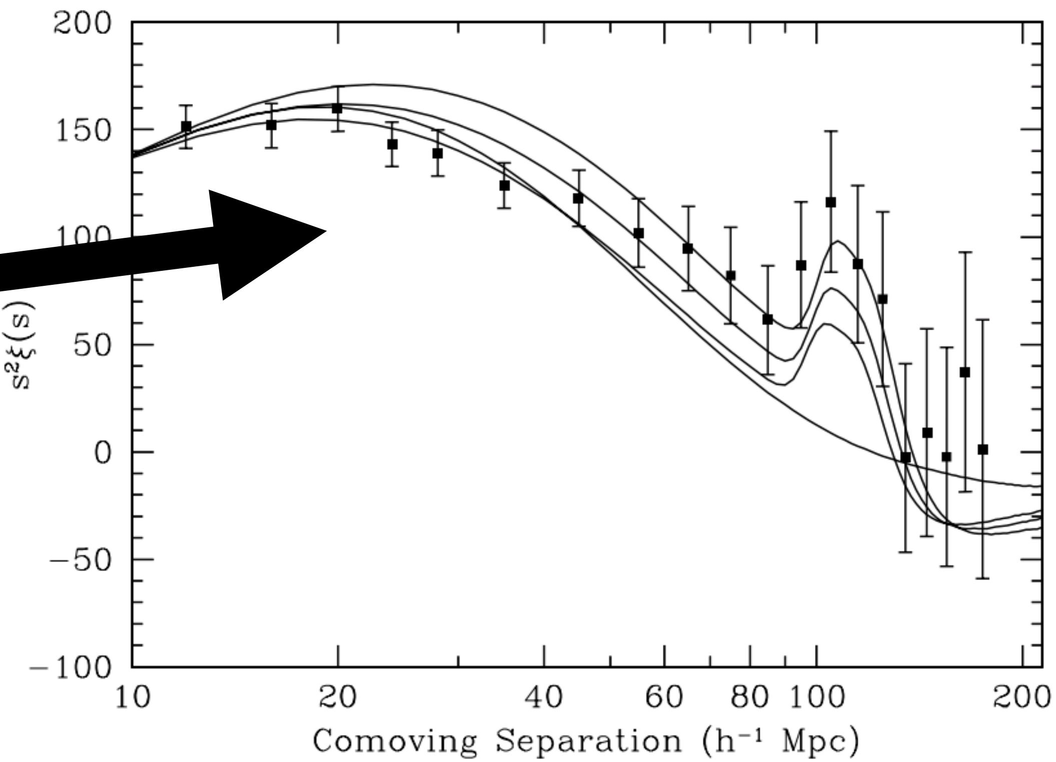
NASA, ESA, and A. Riess (STScI)

type Ia supernovae
(other supernovae)
(quasars)

Probes of the Cosmic Expansion History: Standard Rulers



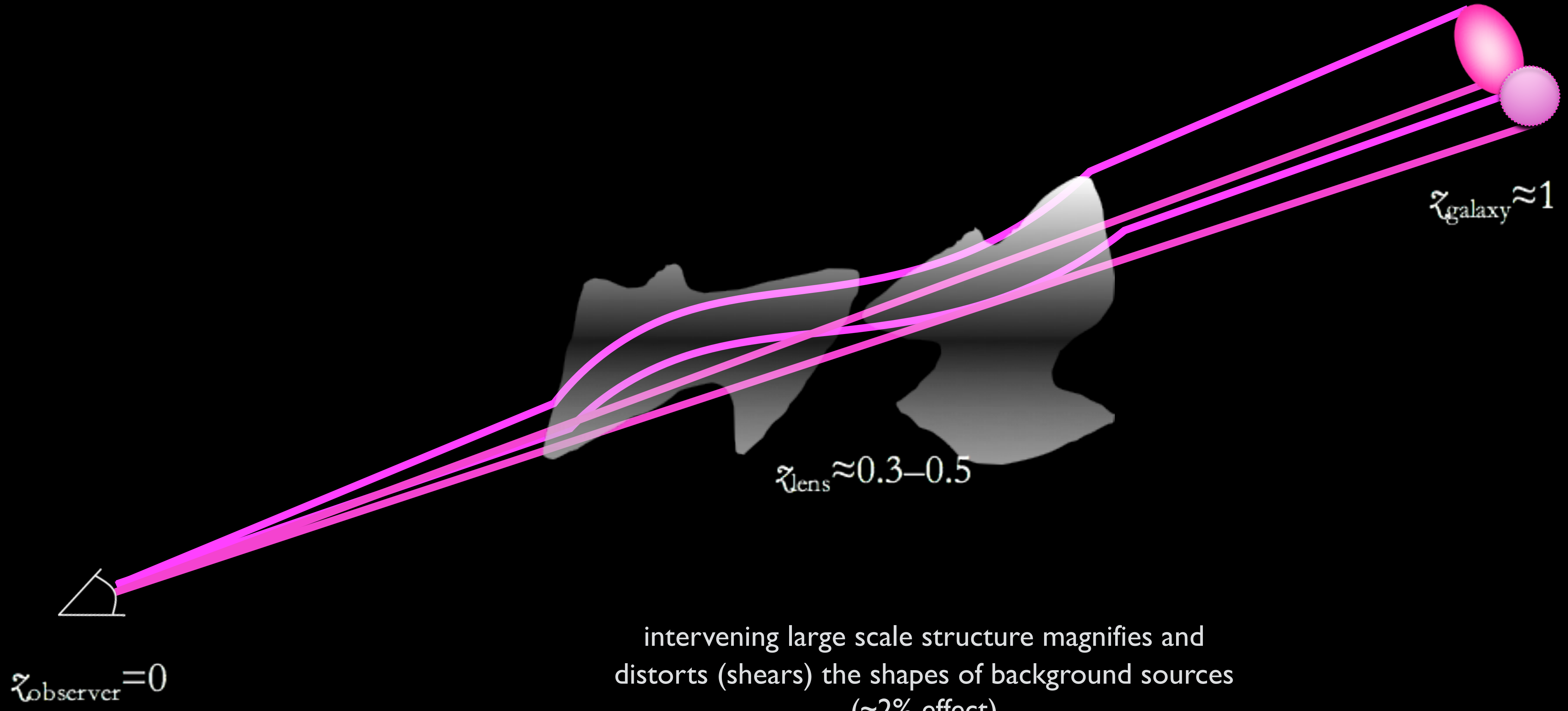
cosmic microwave background



Baryon acoustic oscillations (BAO)

characteristic size is imprinted into cosmic density fluctuations at recombination; can measure that characteristic scale over cosmic time

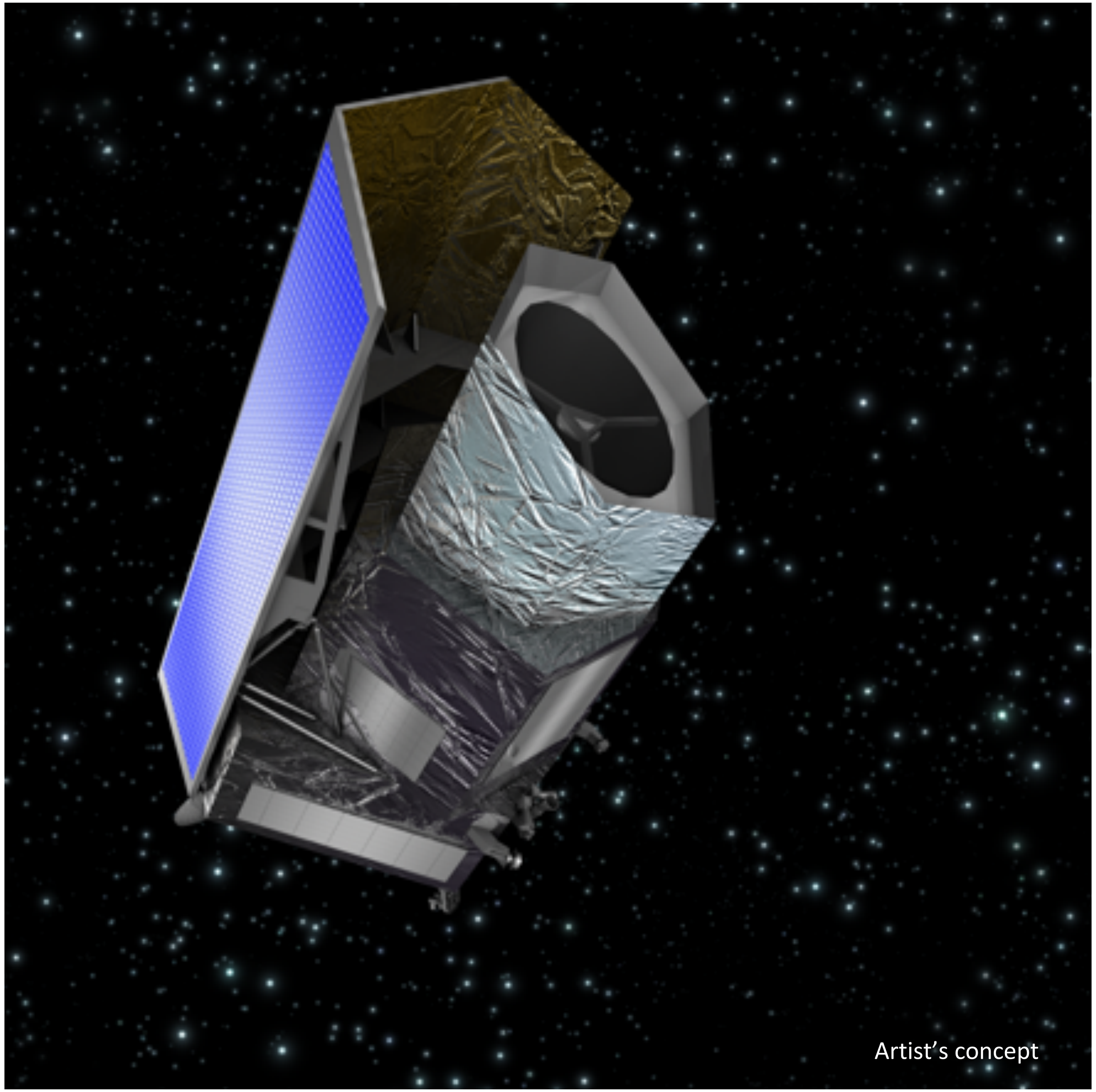
Probes of the Structure Formation: Weak Gravitational Lensing



intervening large scale structure magnifies and distorts (shears) the shapes of background sources (~2% effect)

Upcoming Dark Energy Missions

Proposed lifetime	2022 - 2032	2022 - 2028	2025 - 2031
Mirror size (m)	6.5 (effective diameter)	1.2	2.4
Survey size (sq deg)	20,000	15,000	2,227
Median z (WL)	0.9	0.9	1.2
Depth (AB mag)	~27.5	~24.5	~27
FoV (sq deg)	9.6	0.5 (Vis) 0.5 (NIR)	0.28



Euclid:

- ESA M-Class Mission (~probe-class)
- significant NASA participation
- launch date: **2021**

- 1.2-meter mirror (TMA = three-mirror anastigmat)
- two instruments: **VIS** & **NISP**
- wide-field optical imaging survey (**VIS**)
    - single broad band (riz)
- wide-field near-IR imaging survey (**NISP**)
    - three band (approx. Y, J and H)
- wide-field near-IR spectroscopy survey (**NISP**)

- primary science: cosmology (multiple probes)
- significant legacy science, ranging from resolved stellar populations within ~5 Mpc to most distant quasars

- 6-yr. survey, mapping 15,000 deg^2 from L2

WFIRST:

- NASA Flagship mission
- launch date: mid-2020s*

- 2.4-meter primary mirror (like Hubble)
- **Coronograph + Wide Field Imager / Slitless Spectrograph (0.28 deg^2 FOV)**
- Wide-field imaging / low-res spectroscopy from 0.7-2 μm

- Primary science: cosmology and exoplanets
- Significant legacy science, including early universe galaxies, galactic streams, "extreme" galaxies and quasars, clusters, etc.

- Plan to map ~2,000 deg^2 from L2 for weak lensing cosmology

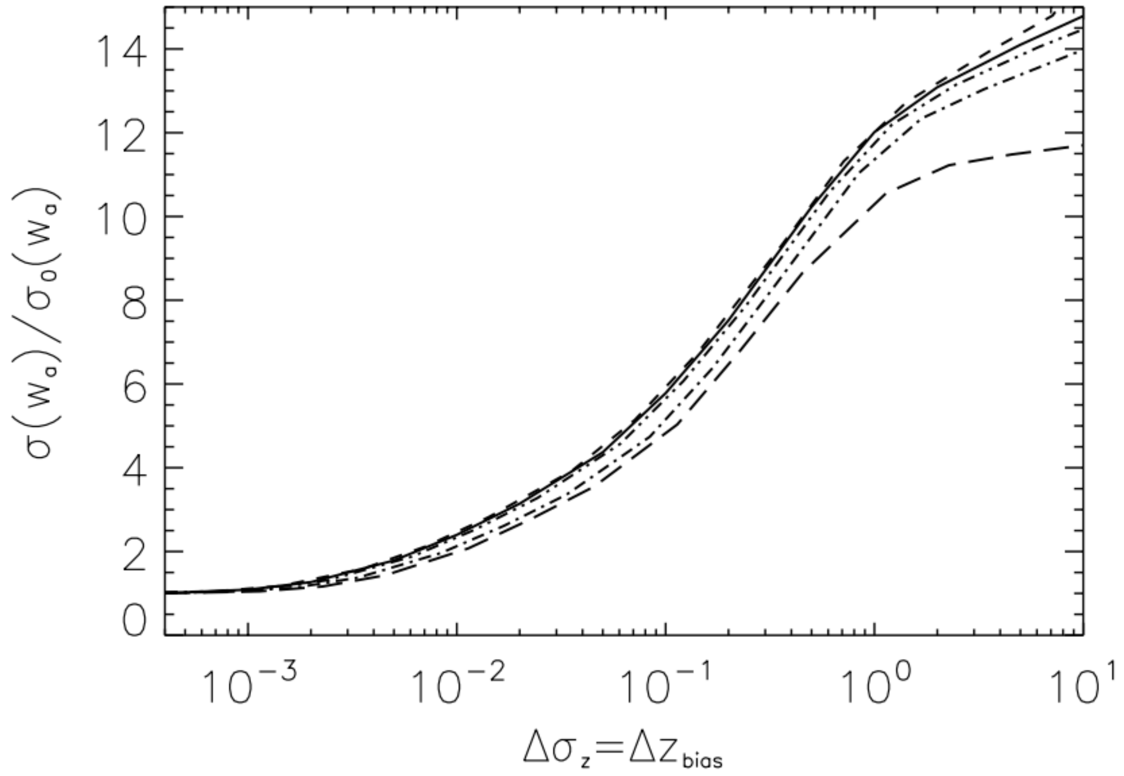
(*Uncertainty is large)

Redshifts for weak lensing cosmology

- Weak lensing probes the growth of structure with redshift
- Need to split shear sample up into well-defined redshift bins, and know the *N*(z) of the galaxies in those bins with high accuracy

➢ Δ⟨z⟩ < 0.002 (1+⟨z⟩) for the redshift bins – i.e., the mean redshift in ~10-20 shear bins must be known to better than 0.2%

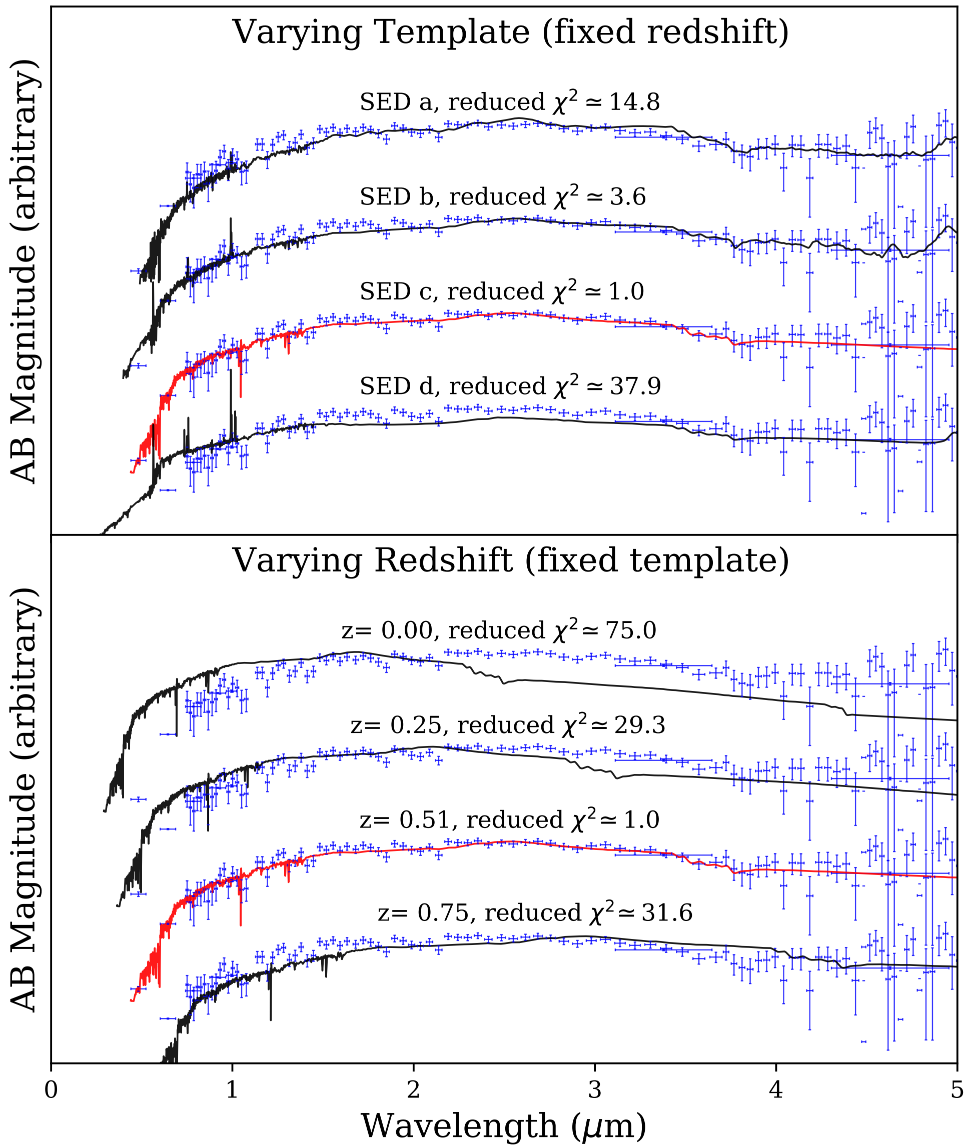
- *Not possible with existing photo-z methods*



Degradation of cosmological constraint with increasing photo-z bias (Ma et al. 2006)

Template fitting

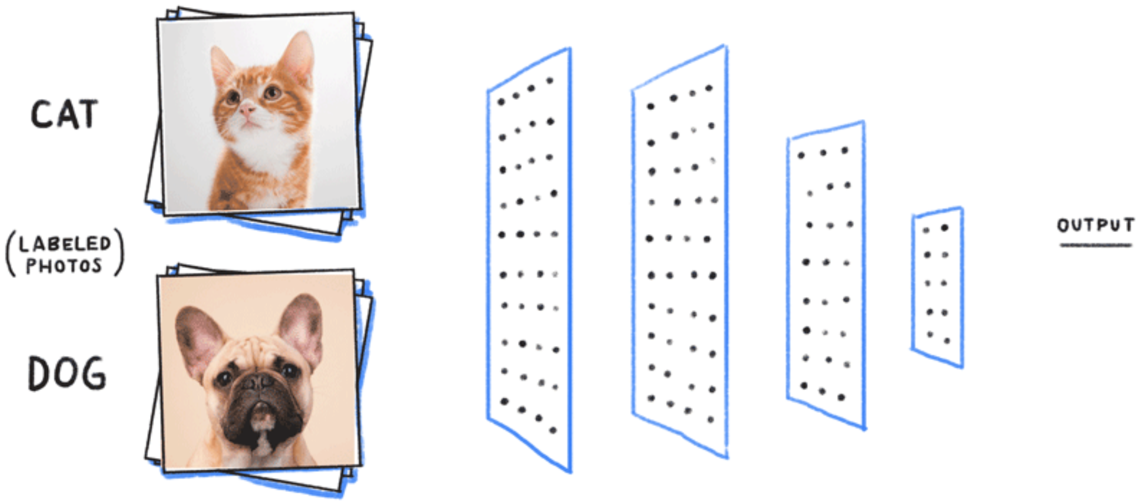
- Use models of galaxy SEDs to define a grid of possible colors
  - Vary redshift, template, E(B-V), reddening law, possibly emission lines, etc.
  - Interpolate against filter profiles to get predicted colors for each permutation
- For observed photometry, use this grid to find the best-fit redshift as well as zPDF
- Main issues:
  - Are templates *fully* representative of the true population? What about overfitting?
  - How to determine correct priors for different template/redshift combinations?



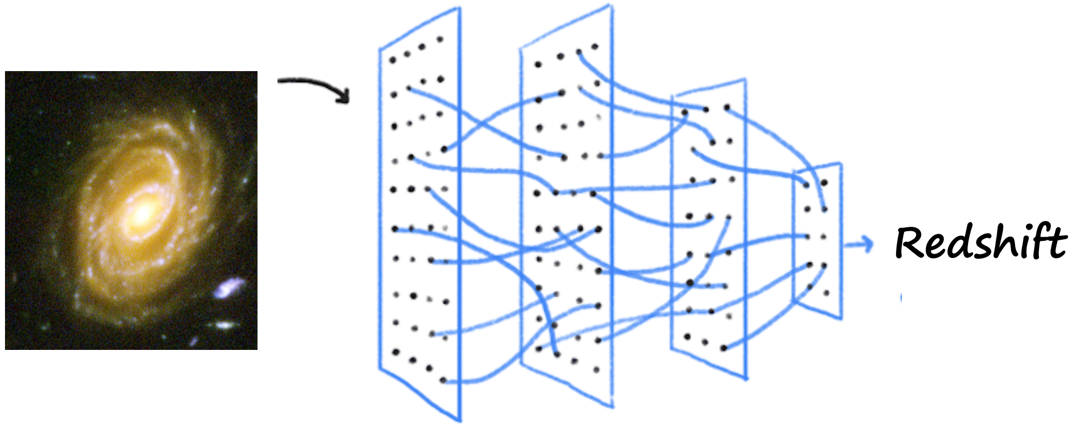
Stickley et al. 2016, simulated SPHEREx data

Machine learning

- Aims to uncover the color-redshift relation directly
- Relies on spectroscopic training samples
- Unfortunately we're *not* in a data rich environment – spectroscopic samples are *limited* and *biased*
  - State of the art ML techniques may not be enough
  - There is no magic solution to biased training samples

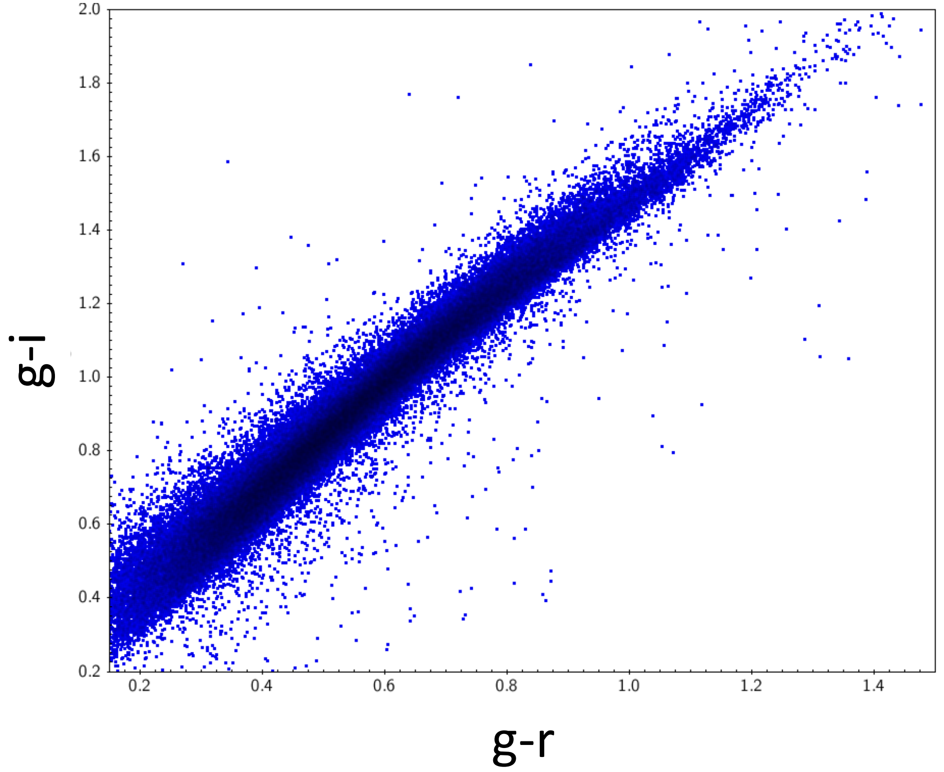


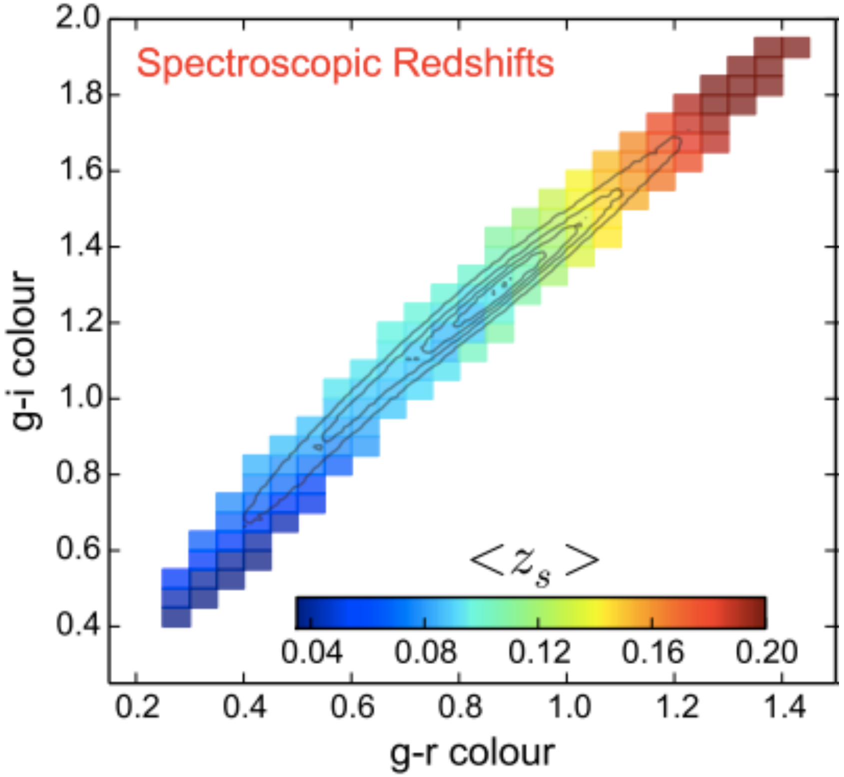
Credit: Google
www.google.com/about/main/machine-learning-qa/



The empirical P(z|C) relation

SDSS galaxy distribution in two colors





Rahman et al. 2015

→ Photo-z's are fundamentally a mapping of galaxy colors to redshift

→ Color distribution of galaxies to a given depth is *limited* and *measurable*

Unsupervised learning approach

- Before we try to understand $P(z|C)$, let's first understand $\rho(C)$ for our survey
  - Map the high-dimensional distribution of galaxy colors
  - Use Euclid-like imaging data from existing deep fields like COSMOS
- Lots of advantages to doing this
  - Can explicitly understand what parts of color space are calibrated with spectra
  - Understand correlations / degeneracies in the data
  - Identify likely outliers based on photometry alone

The self-organizing map

- The problem of mapping a high-dimensional dataset arises in many fields, and many techniques have been developed
- We used the Self-Organizing Map (SOM; Kohonen 1990)

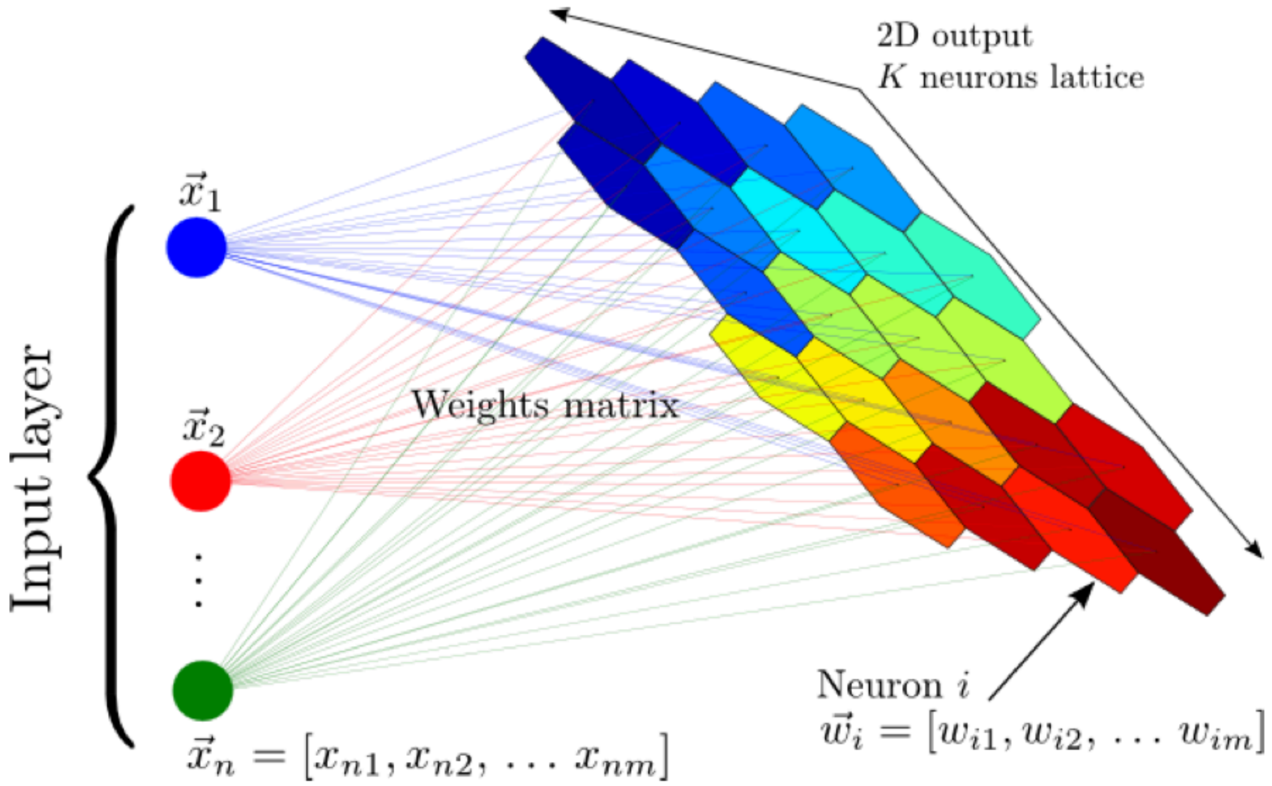


Illustration of the SOM (From Carrasco Kind & Brunner 2014)

What is a SOM?

Starts with high-dimensional data

Credit: Shoubaneh Hemmati (JPL/Caltech)

What is a SOM?

Similar in one dimension

Credit: Shoubaneh Hemmati (JPL/Caltech)

What is a SOM?

Similar in another dimension

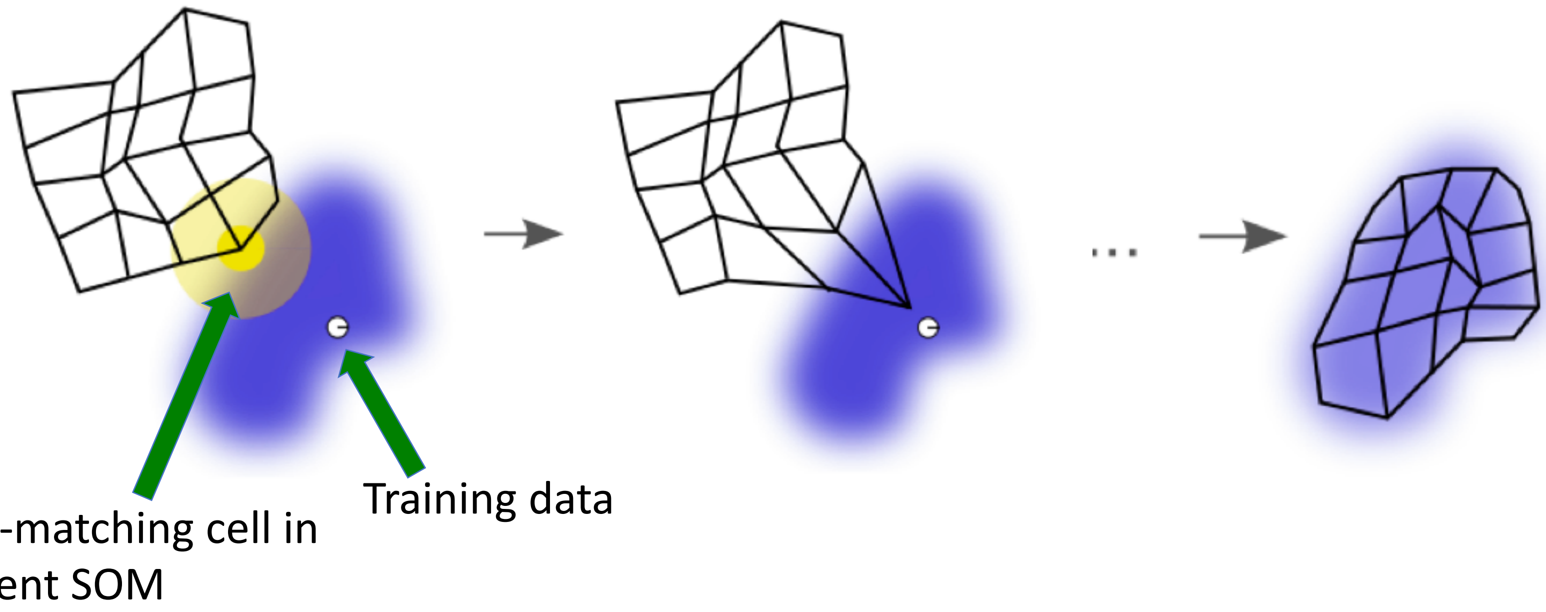
Credit: Shoubaneh Hemmati (JPL/Caltech)

What is a SOM?

- The SOM represents a high-dimensional data space in a topological way. Objects in similar parts of the high-dimensional space are grouped together in the low-dimensional representation.

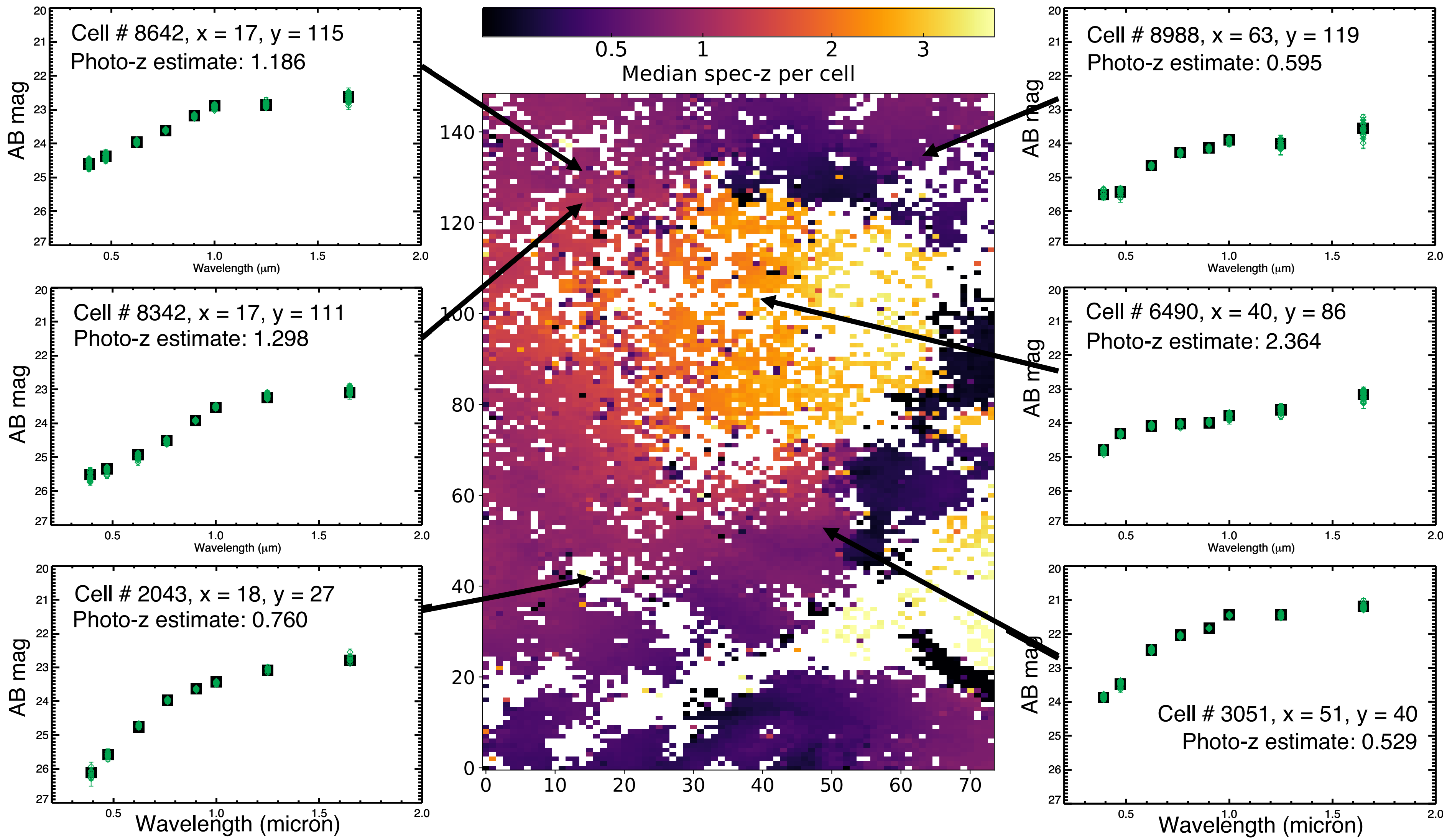
Credit: Shoubaneh Hemmati (JPL/Caltech)

Training the map



1. Initialized map is presented with training data, i.e. the colors of one galaxy from the overall sample.
2. Map moves towards training data, with the closest cells being most affected.
3. Process repeats many times with samples drawn from training set until the map approximates the data distribution well.

The 8-color SOM to Euclid depth with VVDS/COSMOS



C3R2 = Complete Calibration of the Color-Redshift Relation

Judith Cohen (Caltech) - PI of Caltech Keck C3R2 allocation
16 nights (DEIMOS + LRIS + MOSFIRE, kicked off program in 2016A)
Daniel Stern (JPL) - PI of NASA Keck C3R2 allocation
10 nights (all DEIMOS; "Key Strategic Mission Support")
Daniel Masters (JPL) – PI of NASA Keck C3R2 allocation 2018A/B
10 nights (5 each LRIS/MOSFIRE; "Key Strategic Mission Support")
Dave Sanders (IfA) - PI of Univ. of Hawaii Keck C3R2 allocation
6 nights (all DEIMOS) + H20
Bahram Mobasher (UC-Riverside) - PI of UC Keck C3R2 allocation
2.5 nights (all DEIMOS)

+ time allocations on VLT (PI F. Castander), MMT (PI D. Eisenstein), and GTC (PI C. Guitierrez)
-Coordinating closely with these collaborators for these observations
-Sample drawn from 6 fields totaling ~6 deg^2

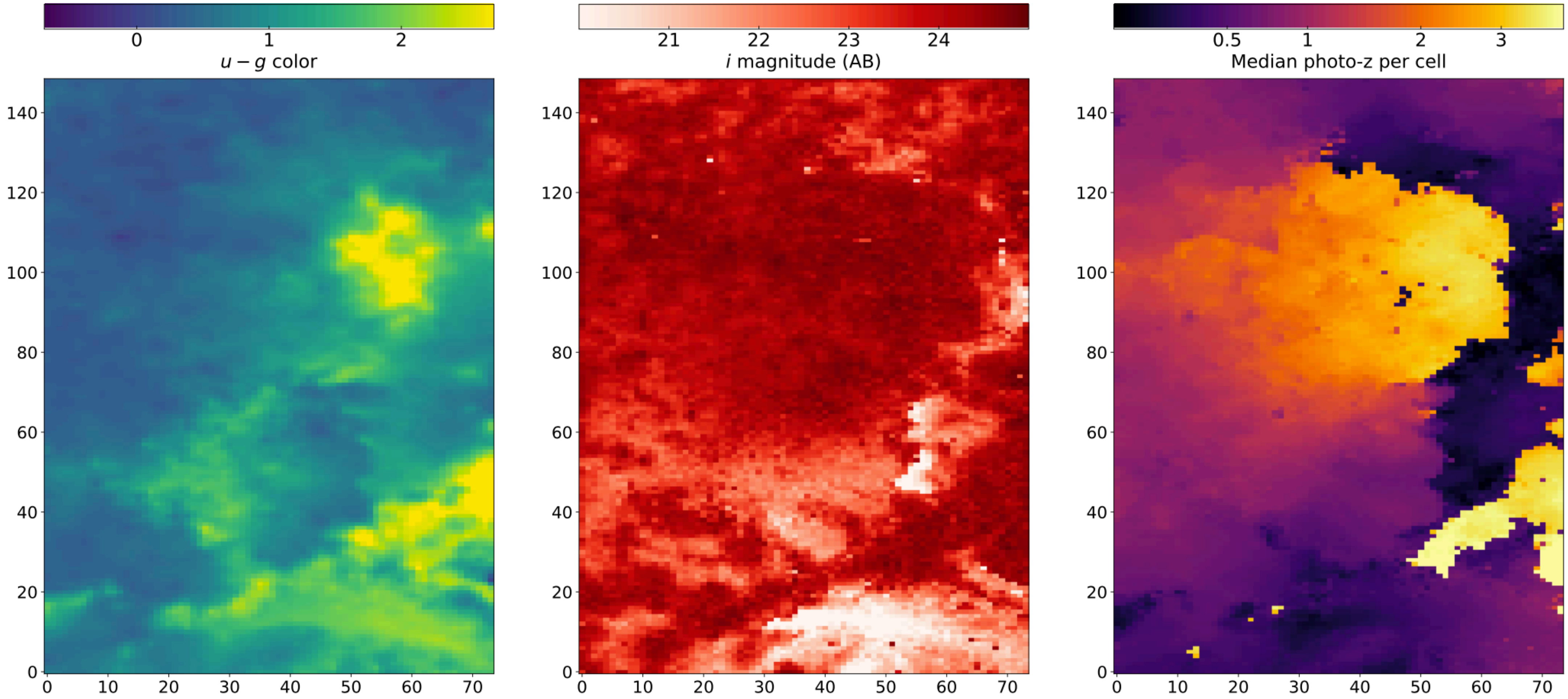
Additional Collaborators: Peter Capak, S. Adam Stanford, Nina Hernitschek, Francisco Castander, Sotiria Fotopoulou, Audrey Galametz, Iary Davidzon, Stephane Paltani, Jason Rhodes, Alessandro Rettura, Istvan Szapudi, and the Euclid Organization Unit – Photometric Redshifts (OU-PHZ) team

Mapping the galaxy P(z|C) relation

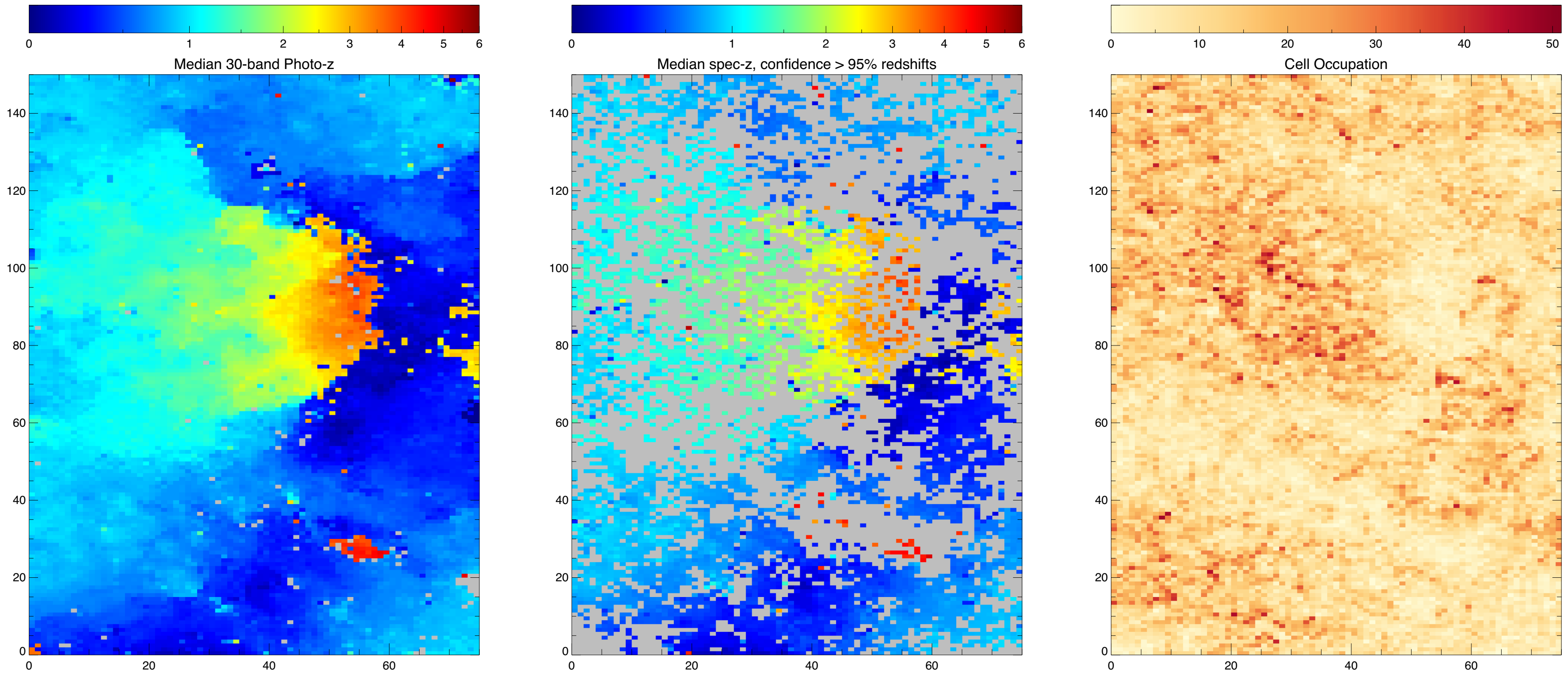
Complete Calibration of the Color-Redshift Relation (C3R2) Survey:

- Designed to “fill the gaps” in our knowledge of the color-redshift relation to Euclid depth
- Collaboration of Caltech (PI J. Cohen, 16 nights), NASA (PI D. Stern, 10 nights, PI D. Masters, 10 nights (2018A/2018B)), the University of Hawaii (PI D. Sanders, 6 nights), and the University of California (PI B. Mobasher, 2.5 nights), European participation with VLT (PI F. Castander)
  - Multiplexed spectroscopy with a combination of Keck DEIMOS, LRIS, and MOSFIRE and VLT FORS2/KMOS targeting VVDS, SXDS, COSMOS, and EGS
  - DR1 (Masters, Stern, Capak et al. 2017) comprised 1283 redshifts, DR2 (Masters, Stern, Cohen, et al. 2019) brings total to >4400 redshifts, observations in 2017B and later will comprise DR3 (**https://sites.google.com/view/c3r2-survey/home**)
- Currently a total of 44 Keck nights awarded (29 observed in 2016A-2017A, 5 nights each in 2017B/2018A/2018B)

C3R2 color map overview



C3R2 survey strategy

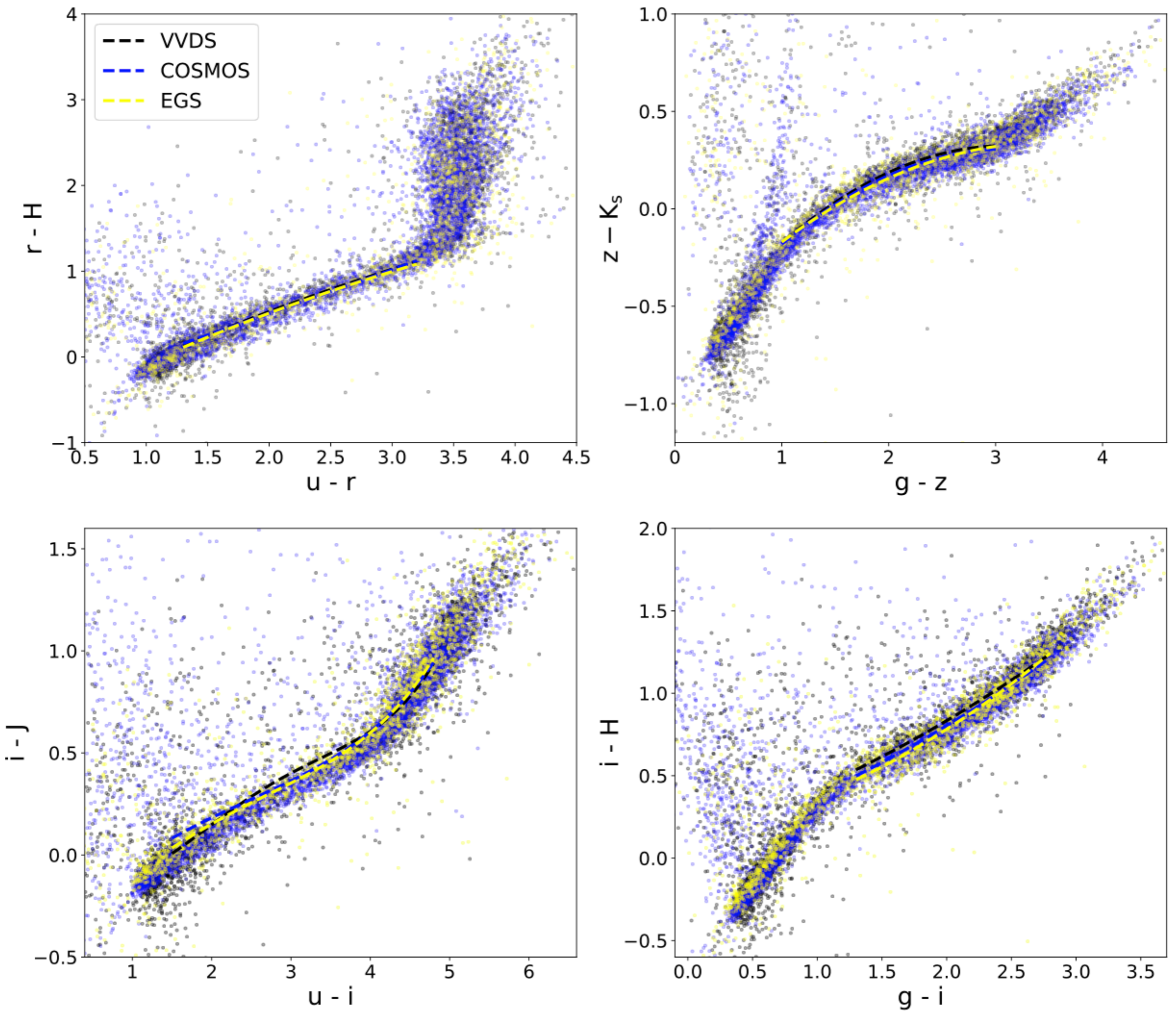


The ingredients of the survey:

Left: Prior on galaxy properties across color space from deep, multiband data
Center: Shows parts of color space that have redshifts and that don’t
Right: Density of sources across color space to Euclid depth

Matched colors across deep fields

- C3R2 targets fields with colors like Euclid at the required depth
- Needed to carefully match the color systems across these fields
- Wound up using CFHTLS photometry in the optical, and combination of CFHT-WIRDS and VISTA data for near-IR



C3R2 stats through DR2 (2016A-2017A)

- 29 nights, ~19 good weather
  - 22 DEIMOS, 5 LRIS, 2 MOSFIRE

- 6696 spectra: 4525 Q >= 3 (high quality), 3970 Q = 4 (certain)

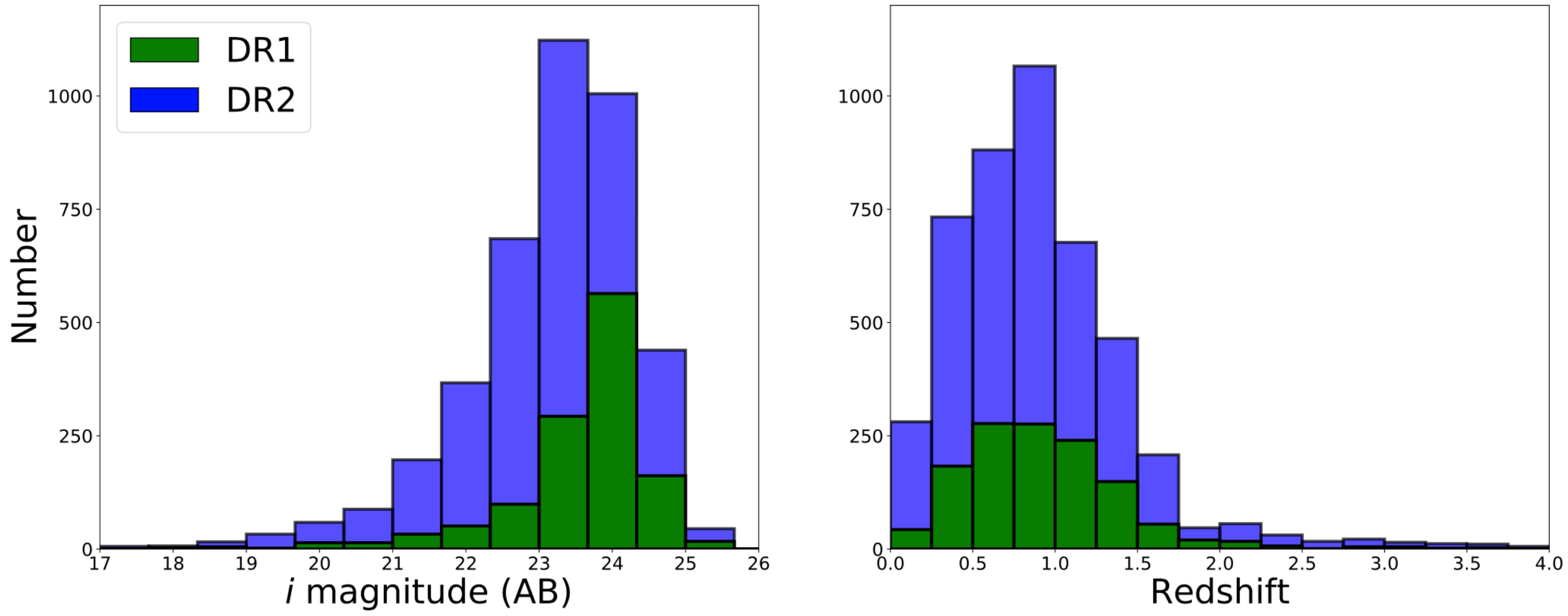
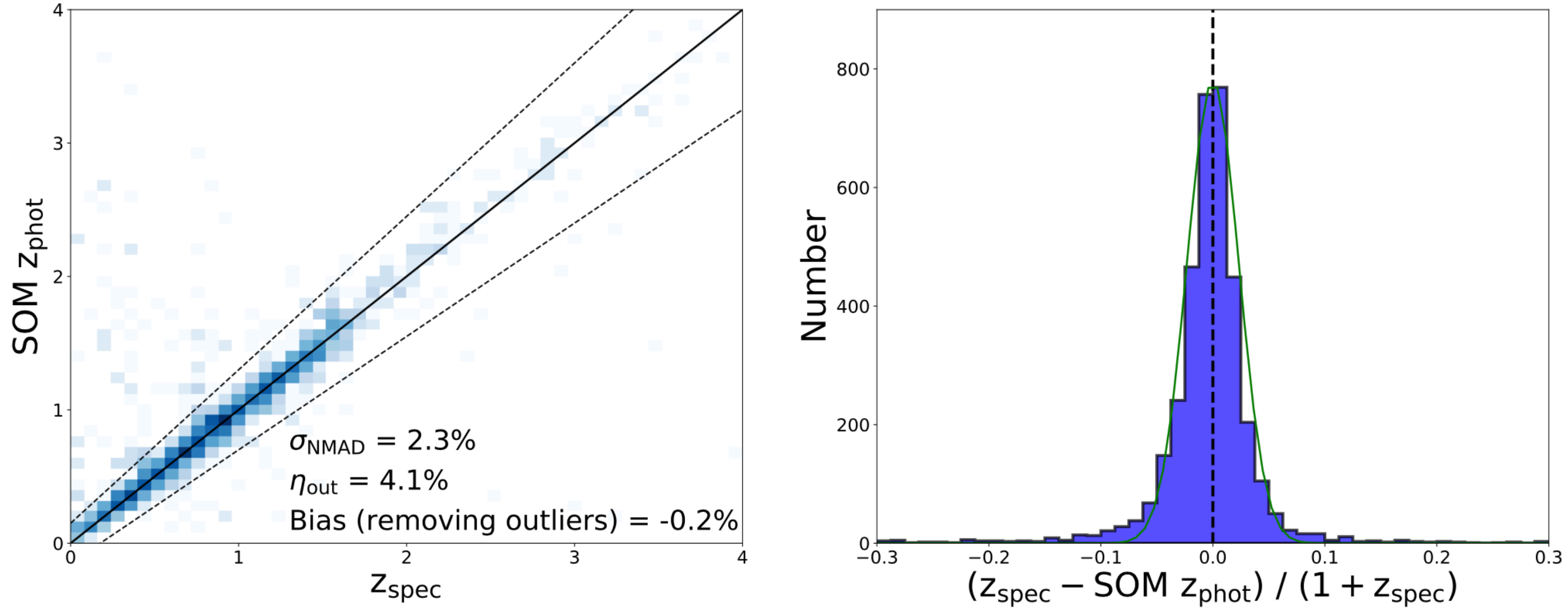


Figure 5. Magnitude and redshift distributions for the C3R2 spectroscopic survey.

Masters et al. 2019 ApJ accepted

SOM-based redshift performance

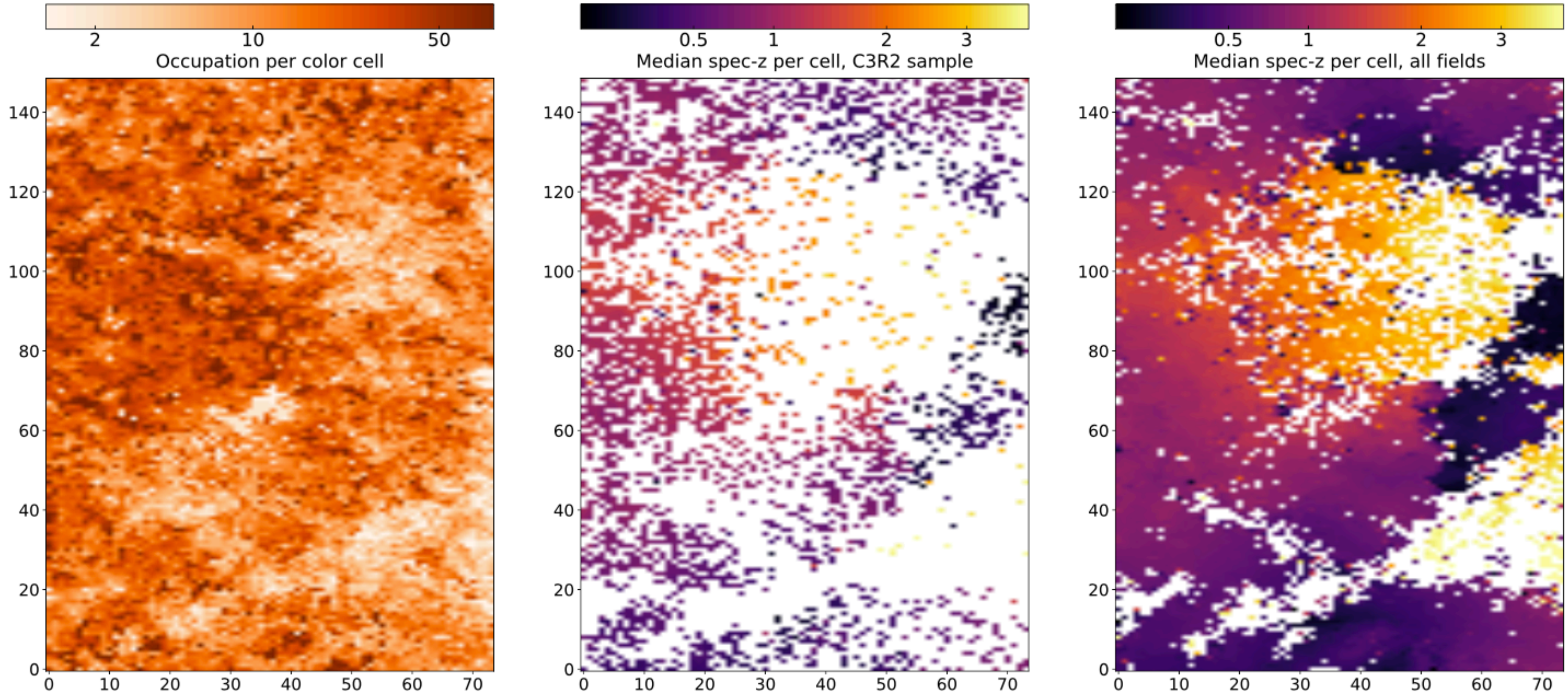
- Simple test: Use position on SOM to predict photo-z
  - Incorporate nothing in defining P(z|C) relation other than the median deep survey photo-z in cells of the Euclid/WFIRST color space



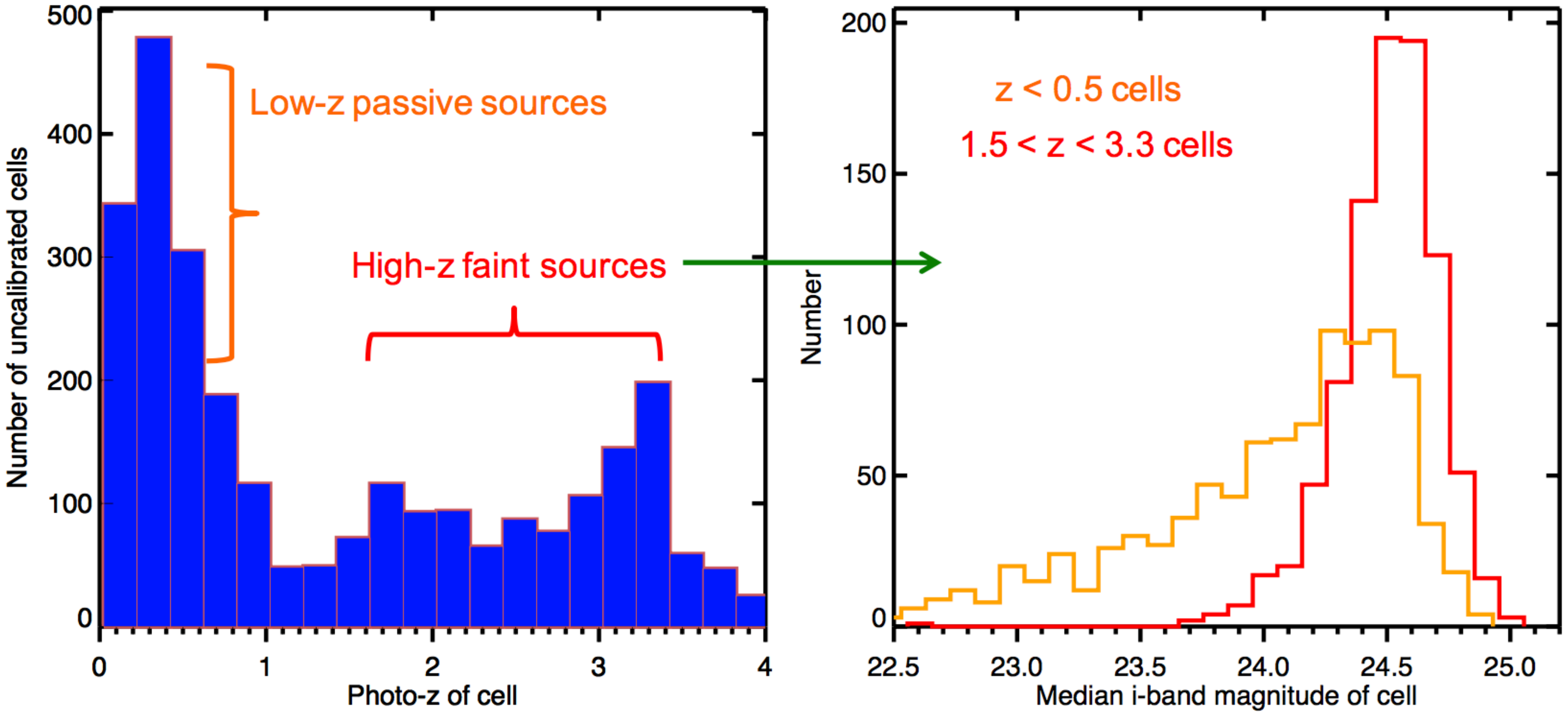
  - Outlier fraction 4.7%, bias (after removing outliers) of 0.18%

Masters et al. 2019 ApJ accepted

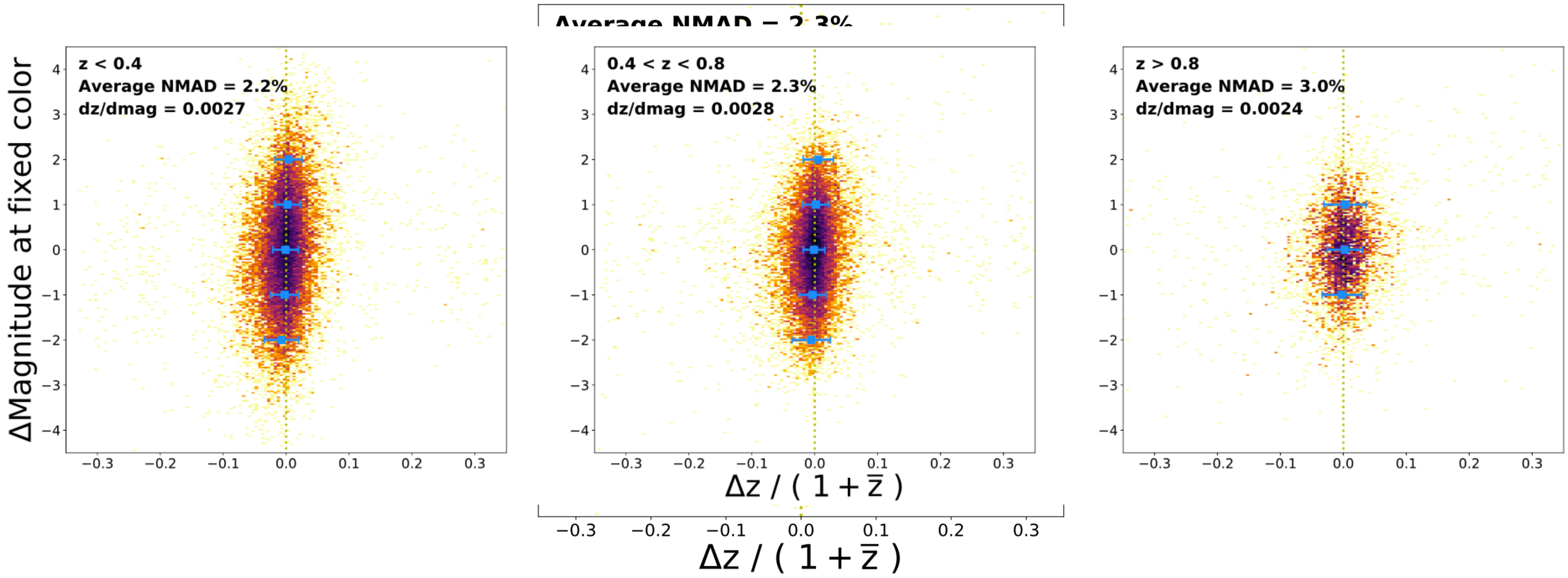
Color space coverage



What are we missing?



How much does galaxy brightness matter?

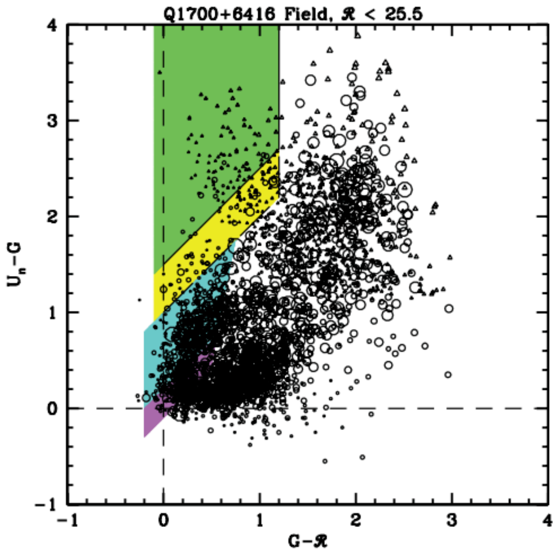


All unique pairs of spec-z galaxies with matching positions on SOM are shown, illustrating the relation of magnitude and redshift at fixed color.

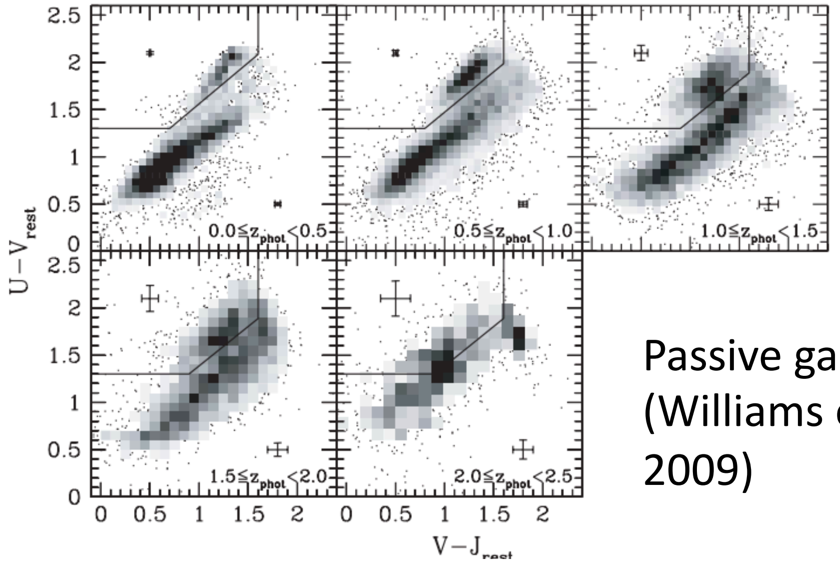
Manifold learning for galaxy physics

- A number of pieces of evidence suggest the information content of the ~8 broadband images is significantly higher than would be inferred from, e.g., template fitting
- We are actively exploring this problem
- Color selections have a long history in astronomy
- What can we learn from high-dimensional color selection?

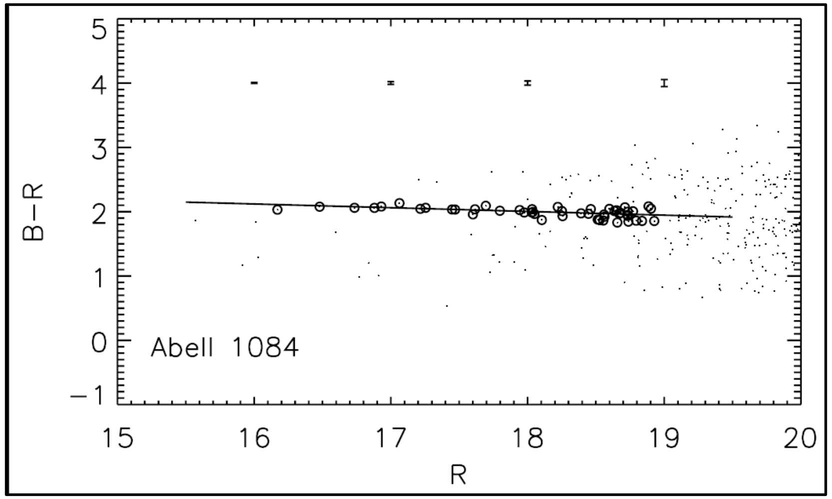
Power of simple color selections



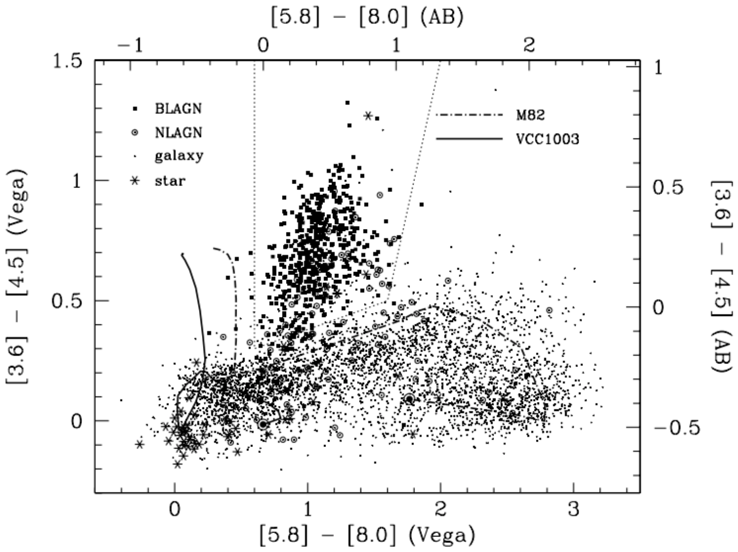
Intermediate redshift galaxies (Steidel et al. 2004)



Passive galaxies (Williams et al. 2009)

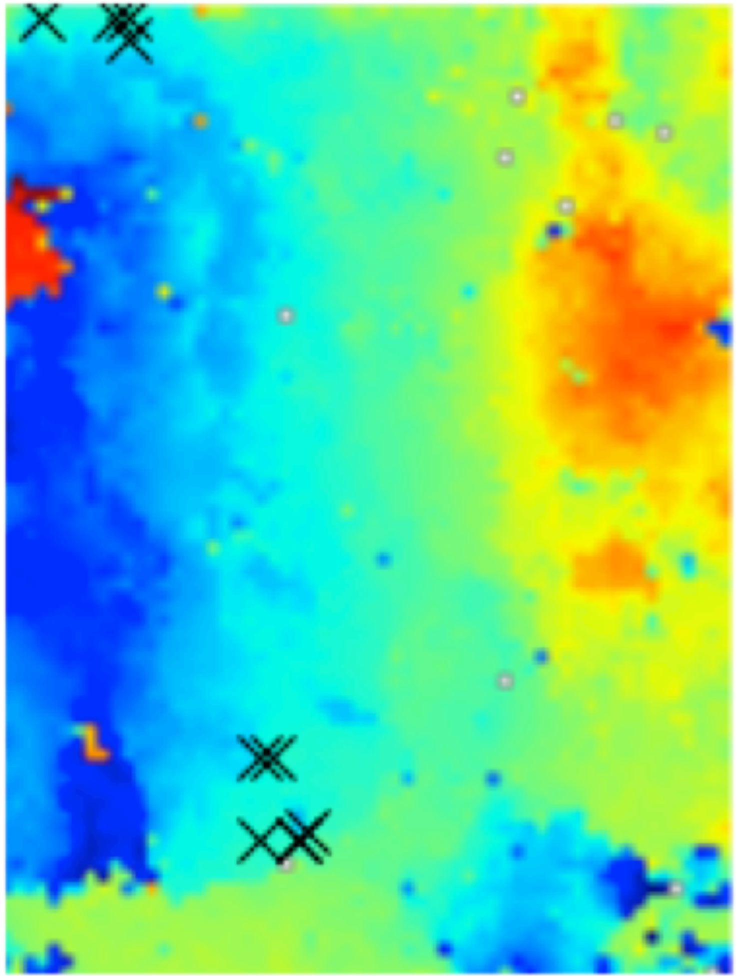


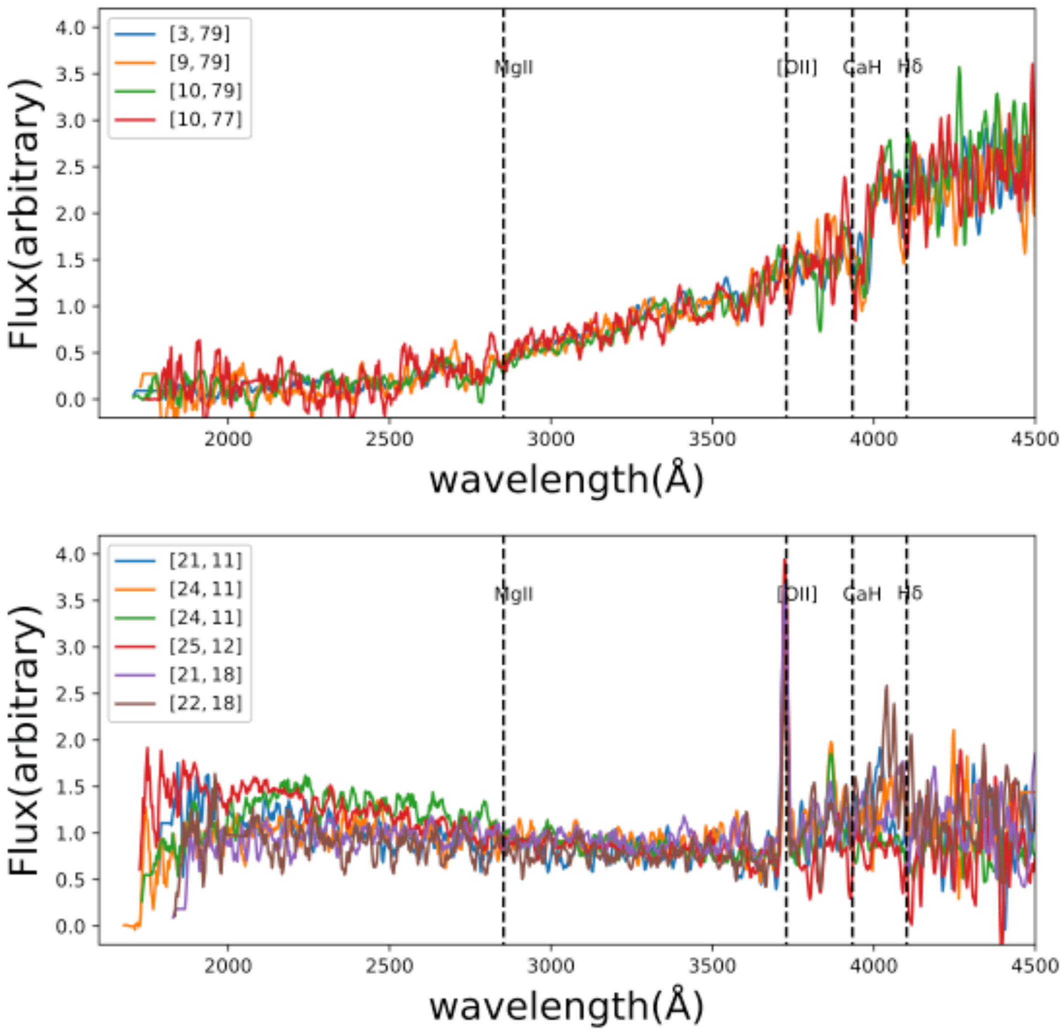
Red sequence cluster selection (Gladders & Yee 2000)



Quasar selection (Stern et al. 2005)

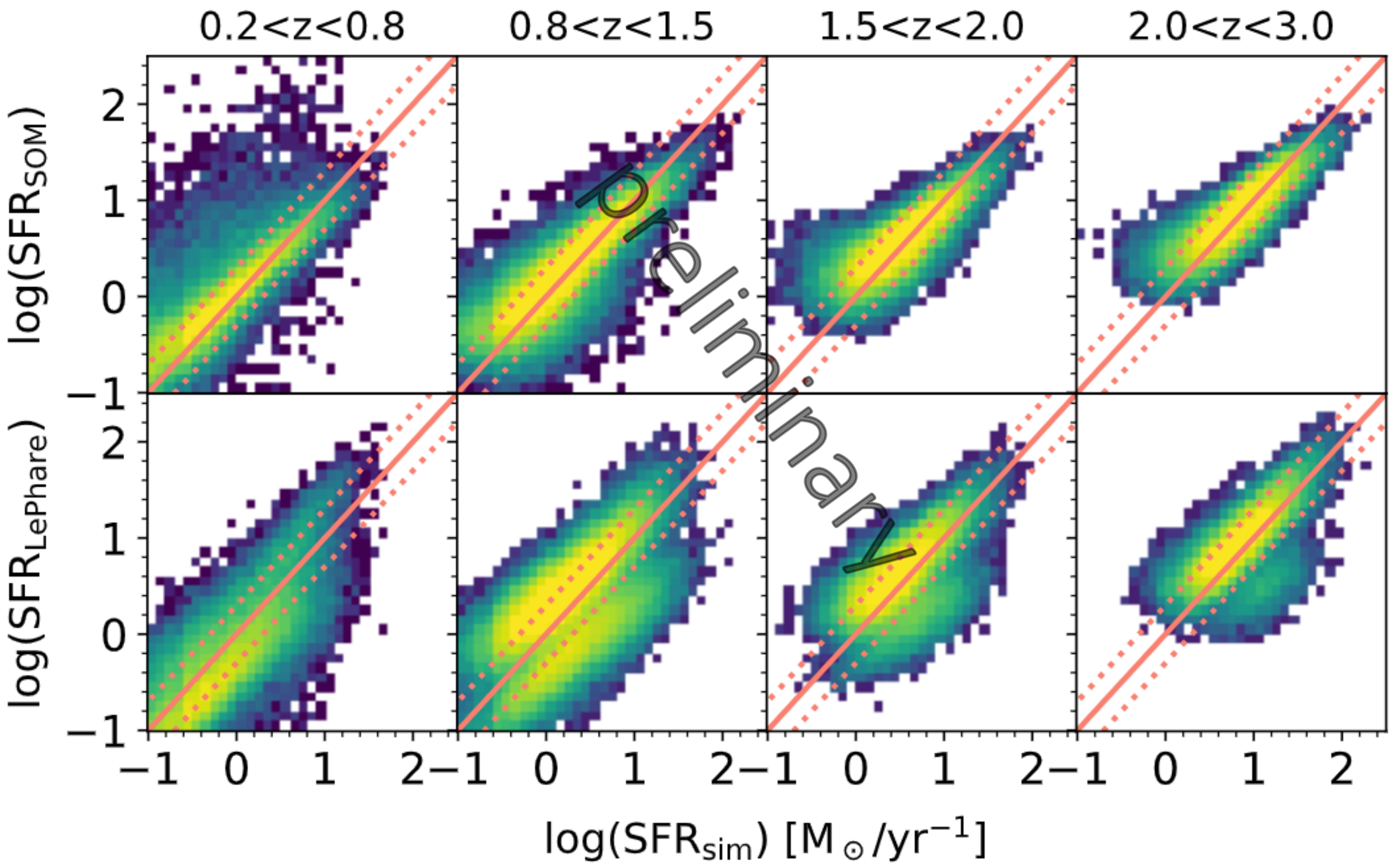
Position on SOM predicts *spectral* properties





Hemmati et al. 2019 (ApJ accepted)

Physics from the manifold



Davidzon et al. in prep

Summary and future research

- Using manifold learning as the basis for redshift calibration for Euclid
- How to complete the color space redshift calibration?
- What are the optimal algorithms for manifold learning / dimensionality reduction?
- Exactly much information is there in the broadband colors we'll have from LSST/Euclid/WFIRST?
- Can we use a C3R2-style approach to better constrain galaxy physics and evolution with these incredibly rich photometric datasets?